Classifying the user intent of web queries using *k*-means clustering

Ashish Kathuria
Department of Electrical Engineering, The Pennsylvania State University, University Park, Pennsylvania, USA

Bernard J. Jansen and Carolyn Hafernik
College of Information Sciences and Technology, The Pennsylvania State University, University Park, Pennsylvania, USA, and

Amanda Spink
Faculty of Information Technology, Queensland University of Technology, Brisbane, Australia

Received 13 July 2009
Revised 5 January 2010
Accepted 16 July 2010

Abstract

Purpose – Web search engines are frequently used by people to locate information on the Internet. However, not all queries have an informational goal. Instead of information, some people may be looking for specific web sites or may wish to conduct transactions with web services. This paper aims to focus on automatically classifying the different user intents behind web queries.

Design/methodology/approach – For the research reported in this paper, 130,000 web search engine queries are categorized as informational, navigational, or transactional using a *k*-means clustering approach based on a variety of query traits.

Findings – The research findings show that more than 75 percent of web queries (clustered into eight classifications) are informational in nature, with about 12 percent each for navigational and transactional. Results also show that web queries fall into eight clusters, six primarily informational, and one each of primarily transactional and navigational.

Research limitations/implications – This study provides an important contribution to web search literature because it provides information about the goals of searchers and a method for automatically classifying the intents of the user queries. Automatic classification of user intent can lead to improved web search engines by tailoring results to specific user needs.

Practical implications – The paper discusses how web search engines can use automatically classified user queries to provide more targeted and relevant results in web searching by implementing a real time classification method as presented in this research.

Originality/value – This research investigates a new application of a method for automatically classifying the intent of user queries. There has been limited research to date on automatically classifying the user intent of web queries, even though the pay-off for web search engines can be quite beneficial.

Keywords User interfaces, Methods of enquiry, Search engines

Paper type Research paper

Internet Research
Vol. 20 No. 5, 2010
pp. 563-581
© Emerald Group Publishing Limited
1066-2243
DOI 10.1108/10662241011084112

1. Introduction

The worldwide web (web) has become a vital tool in many people's daily lives by providing critical access to web resources. Nearly 70 percent of searchers use a search engine as their point of entry to the Internet (Sullivan, 2006). The major search engines receive hundreds of millions of queries per day and present billions of results per week in response to these queries. Search engines are "the tool" that many people use daily for accessing information, Internet sites, services, and other resources on the web.

Prior research has found that searching episodes can be classified in four ways (see Belkin, 1993):

(1) the *goal* of the search interaction;

(2) the *method* of interaction;

(3) the *mode* of retrieval; and

(4) the *type* of resource interacted with during search.

This classification was originally done for earlier forms of searching, such as library systems; however, web searching also possesses these four aspects showing continuity with earlier forms of online searching.

The method of interaction on the web is the same as Belkin (1993) describes for earlier searching systems. The user enters a query, retrieves results, scans results, views results and refines the query as needed. The mode of retrieval is in a hypermedia environment but otherwise remains similar to that of earlier search systems (Marchionini, 1995). However, there are noticeable changes in the goals and types of resources. The goals and types of resources show an example of the long-tail effect of the web, meaning that the web has extended the range of search goals and the range of resources available (Anderson, 2006). Unlike many previous retrieval systems, web search engines provide a variety of alternatives and the resources available are not necessarily uniquely informational, such as transactional services, spelling corrections, and navigation to particular web site.

Web searching is a distinctive domain of study as it differs from earlier searching forms in four respects: context, ubiquity, scale and variety. The first difference between earlier forms of search and web searching is that the range of web content available is larger. Web search engines provide access to both textual and multimedia content. Second, this content can be accessed in almost any setting, whether it is a home, work, or mobile situation. The third difference is the number of searchers. The number of people attempting to find content via web search engines is larger by many times than the number of searchers attempting to access content by older forms of search. Additionally, the range of topics that searchers submit to web search engines is larger than in earlier searching. Lastly, the variety of content, users and systems for web searching is unique. These four points (context, ubiquity, scale and variety) combine to show how extreme the diversity is on the web.

In response, web search engines provide a variety of services for users. The most obvious one is that they attempt to satisfy information needs. They can also serve as navigational tools that take users to specific web sites they want to access by providing links to the specific uniform resource locators (URLs). This allows search engines to aid users in browsing. In addition to navigational tools, search engines can be used to conduct ecommerce transactions. Search engines provide access to other transaction services. For instance, a user can use a search engine to get a map, driving instruction to a specific location, or to find another search engine. Search engines also perform social networking functions (e.g. as on Yahoo! Answers or Mahalo Answers). Web search engines can be spell checkers, thesauruses, dictionaries or entertainment (e.g. Google Whacking, vanity searching). Lastly, search engines can aid in real time and social search by providing results to micro-communication sites such as Twitter.

From the above discussion, modern web search engines have a wide range of features, and they continue to add more. In response to the variety of services and the diversity of content provided, users are continuing to employ web search engines in new, novel and increasingly assorted ways. This increases the necessity of search engines understanding the goal or intent behind a search query in order to provide more useful results to a searcher. Web search engines are popular, but how are people using them to accomplish their intended goal? How can we determine what people are actually seeking when they engage a search engine? What task, need, or goal are people trying to address with their web searching? How can we identify these tasks, needs, and goals? These are the questions that motivate our research.

In our research, we refer to the type of resource a searcher desires as indicated in the user's query as the user intent, based on prior research by Broder (2002). Understanding user intent is an on-going research area in a variety of web related fields (see Lu and Hsiao, 2007). With its greater diversity than earlier systems, modern web search engines can better assist people in finding their desired resources by more clearly identifying what is the intent behind the query.

In this research, we develope a methodology to classify user intent in web searching using k-means clustering. We categorize user searches based on intent in terms of the type of content specified by the query and other expressions of user traits – system interaction, and we operationalize these classifications with defining characteristics. We implement these categories in a program that automatically classifies web search engine queries using k-means clustering. We discuss the results of this classification and examine the error rate for the k-means clustering. Finally we discuss using this approach to improve web search engine performance by providing more results in line with searchers' underlying intent.

2. Literature review

The research reported in this paper uses k-means clustering to classify the need of the user in formulating a query. There has been interest in studying the goal behind users' web queries because it can help to improve search engine performance via page ranking, result clustering, advertising, and presentation of results. Classic information retrieval (IR) typically assumes that the need of the user is informational. For the web, the need is not always informational (see Lee *et al.*, 2005). With web searching, researchers have developed different classifications depending on the users' browsing behaviors (Caramel *et al.*, 1992; Marchionini, 1995; Rozanski *et al.*, 2001) or the queries entered. According to web studies (e.g. Broder, 2002; Jansen *et al.*, 2008; Rose and Levinson, 2004), goals of web searchers are multifaceted, which suggests that the web searcher may be interested in locating a particular web site or a particular product.

Broder (2002) proposed a classification that forms the basis of this research. Broder focused directly on need behind the query rather than analyzing the users' web based activity. He classified the web queries into three categories: navigational (the intent is to reach a particular site), informational (the intent is to acquire information assumed to be present on one or more web pages), and transactional (the intent is to perform some web-mediated transaction). So, user intent is focused on the type of content that the user desires. The query classifications were performed manually and each category was not uniquely defined. Rose and Levinson (2004) provided a more in depth analysis of user intent, providing sub-classes for the main categories of informational, navigational, and transactional. Again, the details of their classification were manual and specific characteristics are not clearly presented.

In addition to the work on manual classification of user intent (Broder, 2002; Rose and Levinson, 2004), other researchers have explored the idea of automatic query classification. Automatically classifying queries allows the classification to be more useful in real time for search engines because the classifications do not require someone to look at and catalog every query by hand. Lee *et al.* (2005) automatically classified informational and navigational queries using a data set of 50 queries collected from computer science students at a US university. They had a 54 percent success rate. Kang and Kim (2003) automatically classified queries as either topic (informational) or homepage related (navigational) and reported a 91 percent success rate. They used documents from a TREC test collection (50 topic and 150 homepage findings) and portions of the WT10 g test collection. Unfortunately, web query classification using retrieved web documents is impractical when dealing with millions of queries (Beitzel *et al.*, 2007), given that one must retrieve and analyze thousands of pages of text and then classify queries all in a fraction of second. Two limitations of the two previous research studies are the small datasets and the limited methodological approaches used. Our research aims to expand both of these so that automatically classifying queries by intent is more useful for search engines in real time.

In automated classification of user intent, Baeza-Yates *et al.* (2006) used supervised and unsupervised learning techniques to classify nearly 65,000 web queries as informational, not informational, or ambiguous. They achieved an approximately 50 percent success rate after clustering queries and did not attempt to classify navigational or transactional queries. Fujii (2008) presents a method for identify navigational queries by comparing them to the anchor text in web pages, using 127 informational and 168 navigational queries. The researcher reports that anchor text can be used for query classification. Cao *et al.* (2009) used the set of previous queries from a user session as well as webpage retrieved by these queries to topically classify queries based on taxonomy of web topics.

Jansen *et al.* (2008) provided a comprehensive automated multi-level analysis, giving detailed characteristics of each level. The researchers reported a 74 percent success rate in user intent classification using a decision tree approach. The researchers report that the ability to automatically classify categories makes these classifications meaningful for use by search engines in real time.

Our paper uses the classification from Jansen *et al.* (2008), which is as follows:

(1) *Navigational.* Queries that demonstrate a desire by the user to be taken to the home page of the institution or organization in question.

(2) *Informational.* Queries with the goal of locating information about a particular topic in order to address an informational need of the user.

(3) *Transactional.* Queries with the intent of obtaining something other than the information.

Each of the three user intents (informational, navigational, and transactional) can be further divided into subcategories. However, for the purposes of this research, we classify the queries at the top level of user intent. Encouraged by the automatic classification results obtained by Jansen *et al.* (2008), this paper aims at using data-mining techniques to more accurately automatically classify queries by user intent in order to improve the search engine results. This continues work in automatic classification of user queries, such as that by Özmutlu *et al.* (2006) that focused on topical classification.

3. Research objective

Our research objective is to automatically classify a large set of queries from a web search engine log as informational, navigational, and transactional.

To accomplish this, we encoded the characteristics of informational, navigational, and transactional queries that we identified from prior work to develop an automatic classifier using *k*-means clustering. We executed the program on three portions of a web search engine transaction log.

4. Research methodology

Our research objective of classifying queries into categories led us to investigate the use of prototype methods. Prototype methods (Han and Kamber, 2001; Hastie *et al.*, 2001) use free estimation techniques (i.e. model free). They are not good indicators of the nature of the relationship between a feature and its class or outcome. However, they can be effective as a "black-box" prediction approach.

In prototype methods, the training data are represented by sets of points in a feature space that represent the prototypes. Each of these prototypes has a class label associated with it. Each point *x* (a query in our research) is classified to the class of the closest prototype. Various prototype methods exist (e.g. *k*-means, *k*-center, etc.). The different methods mainly differ in the number and position of the prototypes used.

4.1. K*-means clustering*

We briefly describe the *k*-means algorithm that we use (Alsabti *et al.*, 1998). The *k*-means prototype methodology is well suited to classifying objectives into a predetermined set of clusters or centroids, which is a good algorithmic representation of our data (i.e. queries with attributes). Additionally, one can use the calculated centroid characteristics to classify new data points, which is a technique that we used in this research.

The overall heuristical process is:

(1) Choose *k* cluster centers and initialize them to randomly defined points inside the dataset.

(2) Optimize the dataset by assigning each sample to its closest prototype using their Euclidean distance.

(3) Update the centroids (prototypes) by computing the average of all the samples associated with that prototype.

(4) If the convergence criterion is not met return to step 2. The usual convergence criterion is where the decrease in the objective function is less than a threshold limit.

The objective function is:

$$L(Z,A) = \sum_{i=1}^{N} \|x_i - z_{A(x_i)}\|^2$$

For *k*-means classification, more specifically, the dataset is broken into training and test sets. *K*-means assumes that training samples are tightly clustered around the prototypes. Hence, the prototypes can work as a correct and a compact reflection of the training data set. *K*-means assumes there are m prototypes which are denoted by:

$$Z = \{z1, z2, \ldots\ldots, zm\}$$

Each prototype is assigned to the nearest cluster. The optimal assignment function A (.) is made to follow the nearest neighbor rule, which is:

$$A(x_i) = \arg\min_{J\in\{1,2,\ldots\ldots,M\}} \|x_i - z_j\|$$

If A (.) is fixed then, the prototype z_j should be the average (centroid) of all the samples associated with that prototype j.

$$z_j = \frac{\sum x_i}{N_j}$$

where N_j is the number of samples associated with prototype j.

The goal of the technique is to minimize the total mean squared error between the training samples and their corresponding prototypes, that implies, the trace of the pooled within cluster covariance matrix.

$$\arg\min_{Z,A} \sum_{i-1}^{N} \|x_i - z_{A(x_i)}\|^2$$

Leveraging this approach, we followed the following steps in using the k-means algorithm to classify user intent of queries from our dataset:

(1) Apply *k*-means clustering to the entire training data set using M predefined prototypes.

(2) For each prototype, count the number of samples associated with it. Then assign the class to the prototype that has the highest count (e.g. if the majority of samples associated with a prototype are informational than assign the prototype to the class informational. This means that each prototype is assigned to one class informational, transactional, or navigational.

(3) Classify a new entry *x* to the class of the closest prototype. Continue for the entire data set.

Using the *k*-means approach, one employs training data to determine the centroid's for each of the clusters. Therefore, for the training data, the algorithm is in iterative mode (i.e. one iterates the number of clusters to achieve some threshold) and continue until convergence is achieved. This is the clustering, or unsupervised classification portion. However, for test data, one does not iteration within the algorithm, instead adding new data points to their nearest centroid. Thus, after the training phase is finished, the centroids are fixed points. The final position of the centroids obtained from the training data is then compared to the final position of the centroids obtained from the test data to generate a percentage error.

5. Implementation of *k*-means algorithm

We used a transaction log from Dogpile (www.dogpile.com) for this research. Jansen and Spink (2007) presented a statistical analysis of a Dogpile transaction log and indicate that the user searching characteristics for this Dogpile search log are consistent with the reported observed characteristics from logs of other web search engines (e.g. Park *et al.*, 2005; Silverstein *et al.*, 1999). Thus, we predict that the classification of user intent for other search engines will be similar.

The search log consists of 4,056,375 records of searches collected on 15 May 2006. Each record contains the following fields:

- *User identification*: a user code that the Dogpile server automatically assigned in order to identify a particular computer.
- *Cookie*: an anonymous cookie that the Dogplie server automatically assigned in order to identify unique users on a specific computer.
- *Time of day*: the recorded time as measured in hours, minutes, and seconds by the Dogpile server.
- *Query terms*: the exact terms of the query as entered by a specific user.
- *Source*: the type of content collection the user is searching in (e.g. web, images, audio, or video). web is the default source.

The 4,056,374 records of the flat ASCII transaction log file were imported into a relational database and a unique identifier was generated for each record. The fields of "Time of day", "User identification", "Cookie", and "Query" were used to locate the initial query of a session and then recreate the series of actions in the session. There is no definitive way to identify agent submissions, but we used a similar approach to that used in prior work by having an upper cut-off on queries (Silverstein *et al.*, 1999). We used 100 queries as the cut-off. The approach does not guarantee that all agent sessions were removed. However, it ensured that most sessions by human searchers were included.

In order to minimize skewing the results by result list viewing, we collapsed the search using user identification, cookie, and query. This eliminated duplicates of result page viewing. In addition, all records with null queries were removed. After this pre-processing of the transaction log, the database contained 1,874,397 queries from 666,599 users (identified by unique IP address and cookie) containing 5,455,449 total terms with 4,201,071 total interactions. The interactions included submitting a query, viewing a search engine results page or clicking on a URL. Table I provides the overall statistics for the dataset.

We calculated three additional attributes for each record:

(1) *Query length*: the number of terms contained in a particular query.

(2) *Results page*: a number representing the search engine results page (SERP) viewed (blank is first page, 1 is second page, etc.) during a given interaction.

Table I.
Dogpile transaction log aggregate statistics

Category	Number	Percent
Users	666,599	
Queries	1,874,397	
Total interactions (queries, page views, and click-throughs)	4,201,071	
Terms		
Unique	360,174	6.6
Total	5,455,449	
Mean terms per query	2.83	
Terms per query		
1 term	352,285	52.8
2 terms	114,391	17.2
3 + terms	199,923	30.0
	666,599	100.0
Users modifying queries	314,314	47.15
Repeat queries (submitted more than once by two or more searchers)	152,771	11.6
	1,159,764	88.4
Unique queries (submitted only once in the entire data set)	1,312,535	100.0
Session size		
1 query	352,285	52.8
2 queries	114,391	17.2
3 + queries	199,923	30.0
	666,599	100.0

(3) *Query reformulation*: the number of times a user changed the query during a session. We used the algorithm outlined in Jansen *et al.* (2007) to classify the queries. An initial query for a session was identified by the IP address and the unique cookie. Subsequent queries were seen as reformulations or changes of the query. A new session started when a query had no terms in common with the initial query for the session.

The next step in the process was to convert this mainly categorical data to numeric data to apply the data-mining tools. We converted the textual fields into a numeric field (Table II).

Table II.
Assignment of numbers for fields

Source	User intent terms	Query reformulation	Equivalent number
Web	Navigational terms	Original query	1
Video	Navigational domains	Reformulated once	2
Images		Reformulated twice	3
Audio		Reformulated thrice	4
News	Obtain terms		5
	Download terms		6
	Entertainment terms		8
	Co-existing terms		9
	Single terms		10
	File extensions		11

The assignment of terms as informational, navigational, or transactional was based on the process outlined in Jansen *et al.* (2008), which we briefly present here. Queries characteristics for each category were:

(1) Navigational searching:
   - queries that contain company/business/organization/people names;
   - queries that contain domains suffixes;
   - queries that have "web" as the source;
   - queries that have a length (i.e. number of terms in query) less than 3; and
   - searcher that views the first search engine results page.

(2) Transactional searching:
   - queries that contain terms related to movies, songs, lyrics, recipes, images, humor, and porn;
   - queries that have "obtaining" terms (e.g. lyrics, recipes, etc.);
   - queries that have "download" terms (e.g. download, software, etc.);
   - queries that relate to image, audio, or video collections;
   - queries that have "audio", "images", or "video" as the source;
   - queries that have "entertainment" terms (pictures, games, etc.);
   - queries that have "interact" terms (e.g. buy, chat, etc.); and
   - queries that have movies, songs, lyrics, images, and multimedia or compression file extensions (jpeg, zip, etc.).

(3) Informational searching:
   - queries that use question words (i.e. "ways to," "how to," "what is", etc.);
   - queries that use natural language terms;
   - queries that contain informational terms (e.g. list, playlist, etc.);
   - queries that were after the first query submitted in a session;
   - queries in which the searcher viewed multiple results pages;
   - queries with a length (i.e. number of terms in a query) greater than 2; and
   - queries which do not meet criteria for navigational or transactional.

There were some navigational queries that were quite easy to identify, particularly those queries containing portions of URLs or even complete URLs. It may seem counter intuitive to some readers, but prior work has noted that many web searchers submit URLs into search boxes as a shortcut to typing the complete URL in the address box of a browser (Jansen *et al.*, 2005). We also labeled company and organizational names as navigational queries, assuming that the user intended to go to the web site of that company or organization. Obviously, there may be other motivations for a user to enter a URL or proper name. We also observed that that navigation queries were generally short in length and occurred at the beginning of the user session.

Our identification of transactional queries was through the user of term and content analysis, specifically via the identification of key terms related to transactional domains such as entertainment and ecommerce.

Informational queries were the default. We did note some characteristics indicating informational searching. The most distinct characteristic was the use of natural language phrases. Informational queries were also longer as measured by number of terms, and the sessions of informational searching were longer in terms of the number of queries submitted.

Next, we converted the query values into meaningful numeric values. This involved converting the string into a vector with each term in the query being converted to a number depending on its category. We based this term classification on prior work outlined in (Jansen *et al.*, 2008). The vector thus formed was converted to an equivalent number by calculating its root mean square (RMS):

$$RMS = \frac{\sqrt{\sum_{i=1}^{n} x_i^2}}{n}$$

where *x* is an individual term in the query and n is the total number of terms. The value obtained for each query was termed the weight of the query. The RMS provided us a measure of the magnitude of each vector, making the vectors comparable. Numeric term identifiers were assigned for navigational and transactional terms (Table II). So, in essence, we pre-processed the log file using a decision tree approach prior to implementing our *k*-means clustering methodology.

Note that for these identifiers within each category the possible terms were broken down into various types or categories. For navigational terms, numbers 1 or 2 were used depending on the level 2 classifications. For transactional terms, numbers 5 to 11 were used again depending on the level 2 classifications. Lastly, the number 0 was assigned to all informational terms.

Because the procedure involved "squaring" the weights, we theorized that if different spectrums for informational, transactional, and navigational terms were managed, results could be achieved with more success. That is, the RMS provided a range for the user intent (e.g. an informational term square would be 0, a navigational term square would be a single digit between 1-4, and a transactional term square would be higher digits between 25-121). So, the RMS for the three categories of user intent each has a different range of values.

After we completed the steps above, the resulting data set had four attributes that we could use for classification: query length, source, query reformulation rate, user intent weight of the query. With the text fields converted to a numeric data, our data set was ready to be processed by the *k*-means clustering technique.

6. Results

We implemented the k-mean data mining technique in three different-sized portions (e.g. 400, 65,000, and 130,000) of the data set. On each occasion, a different number of random queries were chosen for the run. For each run, 75 percent of each dataset was used as training data to estimate the model and the remaining 25 percent was used as test data. We ran three different datasets to provide robustness to our results. The 400 query dataset was used check if the *k*-means model was correct or not, especially to see if the flow adhered to expectations. The 65,000 dataset was used to check the accuracy of the model, and 130,000 dataset was used so that the model was tested against a dataset of large enough size that could be used to draw out general conclusions on the model's accuracy and performance in the real world.

Error percentages were calculated using both the training data and the test data. For the test data, there was no iteration within the algorithm and the final position of the centroids obtained from the training data was used for the test data to calculate the percentage accuracy or agreement between the two data subsets. The data set was classified manually, although snippets of code were utilized to label similar queries. The outputs were recorded for a range of clusters varying from 3 to 20 for each run.

For the first run of the clustering technique (Figure 1) with 400 queries (training set = 250, test set = 150), the lowest error percentage for the training data was approximately 8 percent while that for the test data was approximately 5 percent. The total number of clusters corresponding to this smallest error was 11. The maximum error percentage for the training data was approximately 17 percent while that for test data was approximately 12 percent, indicating an 88 percent accurate classification.

For the second run of the data mining technique (Figure 2) with 65,000 queries (training set $= 48,750$, test set $= 16,250$), the total number of prototypes corresponding to the smallest error was 4 for the training error and 8 for the test error. The maximum error percentage for the training data was approximately 10 percent while that for the test data set was approximately 12 percent (i.e. accuracy rate of 88 percent). As the size of the data sample increases, the test error follows the training error more closely (Figure 2). This trend is more evident in Figure 3.

For the third run of the data mining technique (Figure 3) with 130,000 queries, the total number of prototypes corresponding to this minimum error was 8. The maximum error percentage for the training data set as well as the test data set was approximately 13 percent (i.e. accuracy rate of 87 percent).

Overall, the results indicate that user intent classification by k-means clustering is more accurate by approximately 15 percent than ones obtained by a binary tree classification scheme (Jansen *et al.*, 2008) alone applied to the similar data set. It can

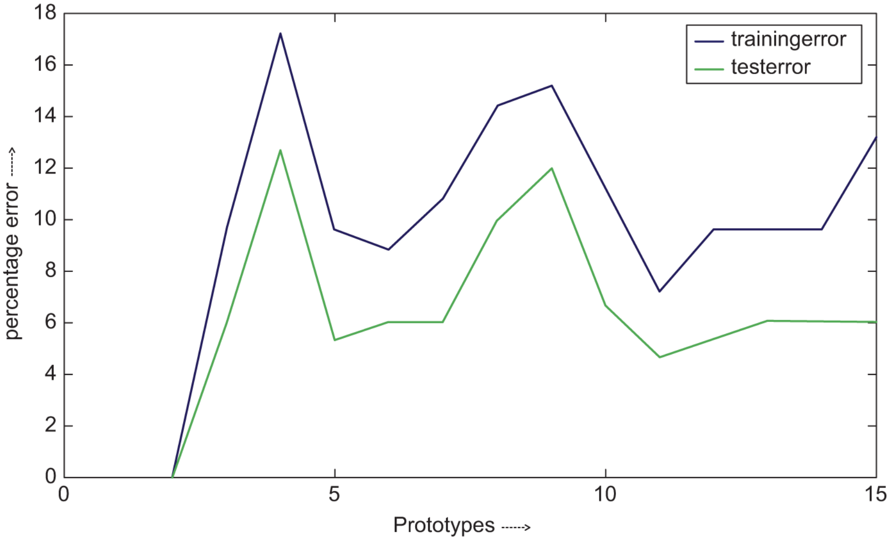

Figure 1.
Error percentage for 400 queries using *k*-means classification

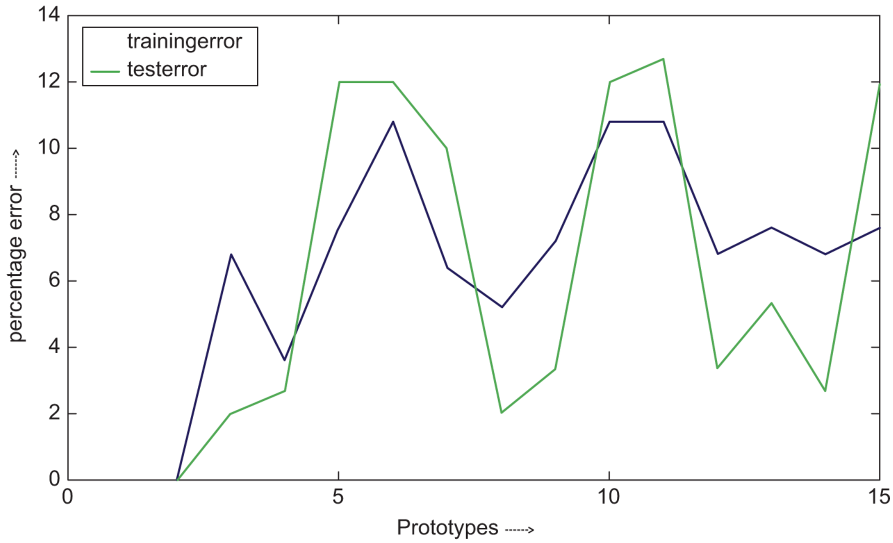

Figure 2.
Error percentage for 65,000 queries using *k*-means classification

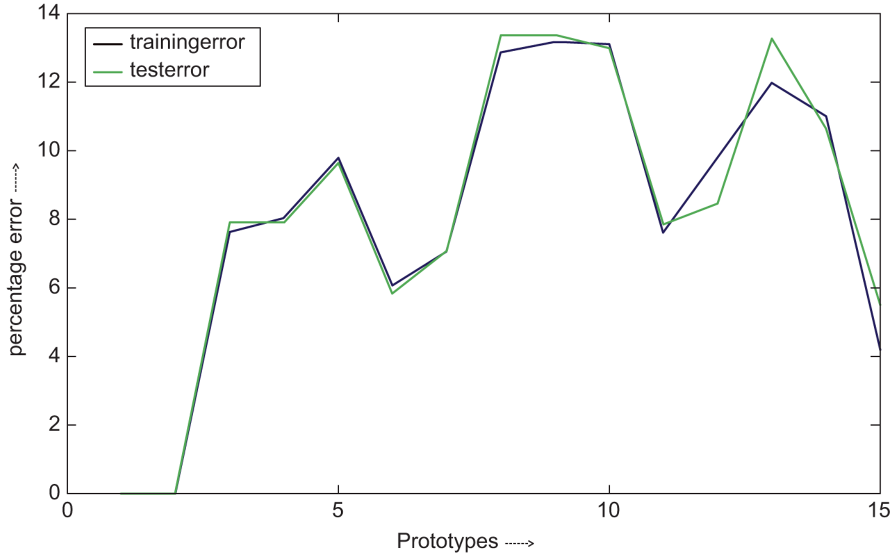

Figure 3.
Error percentage for 130,000 queries using *k*-means classification

hence be concluded that our data mining classification technique is effective for classification of user intent in web queries.

Results indicate that, based on the four features, used in this analysis, that there are eight categories of user intent, calling into the question the practical difference of detailed categories presented in (Jansen *et al.*, 2008; Rose and Levinson, 2004). These other sub-classifications of informational, navigational, and transactional may make logical sense, but they may not exhibit enough unique searching characteristics to permit this automatic classification, if they exist at all.

As seen in Table III, each cluster accounted for about 12 percent of the sample queries. Six of the eight clusters were generally informational, while one was transactional in nature and one was navigational. Overall, about 76 percent of the queries were classified as informational, while about 12 percent were classified as transactional, and 12 percent were classified as navigational. Note that the informational, navigational, and transactional are general labels, representing the preponderance of queries in that cluster.

Some examples of queries classified as informational include “flowering climbing vines” (cluster 6) and “hazards on football fields” (cluster 2). Both of the queries are looking for information about specific topics. They do not seem to be geared towards navigating to a particular site or completing a transaction. Instead, the user appears to wants to find web sites that will provide more information about these two topics.

Navigational queries, on the other hand, include queries where the user is trying to get to a particular web site or class of sites. For instance, two queries classified as navigational (cluster 7) were “university of north carolina” and “image search engines”. One of these queries was attempting to get to a site concerning a university while the other wanted a site where the user could search for images on the web. In both of these cases, the users were trying to locate specific types of sites instead of trying to get information about a subject as the informational queries were.

Transactional queries (cluster 4) were queries attempting to complete a transaction. Examples of this type of query were “download hp simple backup” and “buy camping gear”. For the first query the user appears to have intended to want to find a web site where they could play or download a software program, while in the second query, the user appears to want to purchase camping gear. So, this query has a transactional goal instead of a purely navigational or informational goal.

The differentiation among clusters is presented in Table IV.

As we see in Table IV, for all clusters, the source was primarily web. This would be natural given the predominance of the web as a source is the dataset, relative to

Cluster	Class associated	Percent of test set	Example queries from cluster	Comments on content indicated by queries
1	Informational	12.6	“camper tires”	Specific information needs
2	Informational	12.8	“hazards on football fields”	Open ended informational needs
3	Informational	12.6	“sailor moon”	Specific information needs
4	Transactional	11.5	“buy camping gear”, “download hp simple backup”	Transactional need
5	Informational	12.6	“list of food in alphabetical order”	Informational sites with of lists or catalogs
6	Informational	12.2	“flowering climbing vines”, “historic advent movement”	Open ended informational needs
7	Navigational	12.6	“footlocker”, “craigslist”	Browsing need
8	Informational	13.0	“government grants for small business”	Specific information needs

Table III.
Clusters and classes associated with them

Table IV.
Variations in clusters by attributes

Cluster	Class associated	Percent of test set	Query length	Source	Reformulation
1	Informational	12.6	2.83	Primarily web	2.84
2	Informational	12.8	2.95	Primarily web	**3.07**
3	Informational	12.6	**3.08**	Primarily web	1.70
4	Transactional	*11.5*	2.76	Primarily web	2.81
5	Informational	12.6	2.71	Primarily web	2.99
6	Informational	12.2	2.89	Primarily web	2.29
7	Navigational	12.6	*2.19*	Primarily web	*1.00*
8	Informational	**13.0**	2.98	Primarily web	1.02
Average		12.5	2.80		2.22

Note: The highest numerical value in each column is bolded, and the lowest numerical value in each column is italicized

images, video, audio, or news. Query length varied from just more than two terms to a little more than three terms. What stands out with this characteristic is that the navigational query (cluster 7) has queries that are noticeably shorter than the other classes. This would make sense if the user has a known web site in mind, probably just entering the URL or the web site name in the query box. Transactional queries (cluster 4) and one of the informational groups (cluster 5) had somewhat higher but relatively shorter queries than the other informational classes.

In terms of reformulation, we again see that navigational queries had low rates of reformulation, typically sessions of just one query. Informational clusters 3 and 8 also had low occurrences of query reformulation, indicating probably relatively easy informational needs, such as fact finding. The other clusters had steadily increasing levels of reformulation, with informational cluster number 2 having the highest. Combined with the relatively longer query length, this would indicate that cluster 2 would probably be queries of more complex informational needs.

7. Discussion and implications

This research demonstrates that our approach for classifying queries can be implemented automatically with a high degree of accuracy. Our automated approach using k-means clustering, along with pre-processing using the binary tree approach outlined in (Jansen *et al.*, 2008), achieved a 94 percent success rate, classifying queries into eight unique clusters. When compared to other attempts at automatic classifications, we see that this success rate is quite good. Lee *et al.* (2005) had a 54 percent success rate with 50 queries. Kang and Kim (2003) had a 91 percent success rate but used documents from a TREC test collection, classifying only a couple hundred informational and navigational queries. Baeza-Yates *et al.* (2006) achieved an approximately 50 percent success rate after clustering approximately 65,000 queries. These prior works used much smaller data sets, had higher error rates, and focused only on informational and/or navigational queries.

Our approach not only has a success rate better than those reported in prior work, it also uses a large data set of queries and does not depend on external content, thereby making it implementable in real time. This makes it a viable solution for web search engines attempting to provide relevant content to users. The high success rate means that the classification provided is more reliable and thus more useful for web search engines. The large data set used also makes the approach more realistic for major web search engines as they work with millions of queries per day. Lastly, the fact that our method does not depend on external documents means that web search engines would not need to use external content such as retrieved web pages, and instead could only use the information they already log for each query. Being able to implement our method in real time is important for web search engines because in order for a classification method to be useful, the search engines have to be able to provide results to searches quickly.

In analyzing our findings, there are certain limitations that may restrict the generalizability of our conclusions. One issue is whether the Dogpile user population is representative of web search engine users in general. If it is not, then the resulting classification of query clusters would not be representative of the general web population. Applying our classification methods to data from other major search engines is a goal for future work. It would also be beneficial to do a qualitative analysis of newer transaction logs than the ones used for this study. Such logs might provide increased clarity on characteristics of various user intents. However, Jansen and Spink (2005) report that query characteristics across search engines are fairly consistent. Additionally, Jansen *et al.* (2008) derived their initial characteristics from seven other transaction logs from three other search engines. Therefore, we would expect similar results from other datasets.

Another limitation is that we assigned each prototype to one class of user intent. We are aware that a query may have multiple possible user intents, thus a prototype could have multiple user intents associated with it. Taking this into account, future work will focus on investigating approaches such as naïve Bayes or similar data mining techniques to arrive at a probability of classifying a query into one or more categories instead of using a winner take all approach that produces a binary answer. However, past research (Jansen *et al.*, 2008) suggests that approximately 75 percent of queries can be classified into a single category of user intent (i.e. informational, navigational, or transactional) with a high degree of certainty.

A third limitation of our findings is the inherent shortcoming of relying solely on data from transaction logs. Transaction logs are excellent for collecting large amounts of data from a large number of users engaged in real searching tasks. However, since we do not have access to the users, we can only infer their intent from the data available. We cannot ask users what their actual intent was. An exciting area of future research would be to conduct a laboratory study aimed at gaining further insight into the underlying intent of web searchers. Such a laboratory study would be a good supplement to the transaction log research presented here and in prior work.

There are several strengths of this study, including the variety and size of the dataset employed. Broder (2002) and Rose and Levinson (2004) used a very small number of queries and classified the queries manually. In addition, they did not present their metrics for implementing their classification. Lee *et al.* (2005) used 50 queries, and Kang and Kim (2003) used 200 queries. Baeza-Yates *et al.* (2006) used approximately 65,000 queries but clustered them before categorizing them. Therefore, our results are more robust than the results of previous studies. However, the most important implication of this study is the usefulness of the approach. The approach only uses the characteristics of the current user interaction and query and thus can be implemented for real time classification by search engines. It does not use external information such as already retrieved documents, making the classification quicker and more realistic for large datasets. Thus, search engines can potentially use the information about user intent to provide more relevant results in real time.

Identifying the user intent of web queries could be useful for web search engines because it would allow them to provide more relevant results to searchers and more precisely targeted sponsored links. This would be especially useful in the area of transactional queries. Since transactional queries often carry a higher commercial intent, these queries should be of more interest to online advertisers. For these users, web search engines could more heavily weight results with commercial content or sponsored links. For example, one query classified as transactional was "buy table clock". For this query the search engine might specifically focus on commercial links to retail that allow the user to purchase clocks, which is a continued effort to determine customer preferences (Korgaonkar *et al.*, 2006).

Similarly, targeted actions could be taken for navigational and informational queries. For instance, non-transactional query results could receive fewer sponsored links or sponsored links that highlighted things other than prices or purchasing items. For instance, sponsored links could highlight information that is on or provided by the web site instead of specific purchases. For informational queries, search engines could concentrate on presenting results that are not stores and that do not provide transactional services as their main goal. Navigational queries could focus on presenting results that provided links straight to a requested web page, a person's webpage, or to related web sites. For instance, a common navigational query is "walmart." For a query like this, the search engine might provide links to Walmart, but it could also provide links to other major retailers in the searchers geographical location (using IP look-ups). Another interesting use of navigational queries is the linkage of navigational aspects with informational queries, as raised by (Tann and Sanderson, 2009), who report that users are submitting informational queries with the expectation that a certain web site will be among the results. In other words, these queries have an informational intent and with an element of a navigational expectation.

There are several areas for future research. First, given that there were eight categories of informational queries, a further study investigating these sub-categories would very interesting and potentially beneficial for a search engine to provide more granular informational results. Also, a laboratory study would be a good complement to log analysis. Such a laboratory study might shed further light on how searchers express their underlying intent. Additionally, a detailed qualitative analysis on a search log from another major search engine might lead to more fine grained attributes of user intent across user populations. We would also like to develop algorithmic approaches for utilizing this knowledge of user intent to provide searchers with more targeted results. This is especially noteworthy in the supervised classification area, which would allow for more in-depth analysis of individual queries. We could also explore the use of other approaches than *k*-means clustering, comparing their performance. Finally, we aim to expand our automated classification methods to include the more detailed categories at level two and three (Jansen *et al.*, 2008), which will increase our understanding of user intent and will potentially help search engines provide even more targeted results. A more quantitative understanding of the need behind the query could therefore be a vehicle for producing superior search engines than those in existence.

8. Conclusions and future work

With the increased use of search engines for a varied set of tasks, their effectiveness has become an even more important design aspect. Hence, it is imperative that future research be aimed at improving the performance of search engines across this array of tasks. With this objective in mind, this research aims at improving a search engine's effectiveness by incorporating data mining categorization techniques to automatically classify user entered web queries on the basis of user intent. The use of k-means as an automatic clustering and classification technique yielded positive results and fared much better than the binary tree classification algorithm used previously. The k-means clustering approach to classifying user intent explored in this research has opened up important venues for implementing automatic user intent classification and has considerable potential for future research.

References

Alsabti, K., Ranka, S. and Singh, V. (1998), "An efficient *k*-means clustering algorithm", *Proceedings of the 11th International Parallel Processing Symposium, March 30-April 3*.

Anderson, C. (2006), *The Long Tail: Why the Future of Business is Selling More of Less*, Hyperion, New York, NY.

Baeza-Yates, R., Caldeĺon-Benavides, L. and Gonźalez, C. (2006), "The intention behind web queries", paper presented at the String Processing and Information Retrieval (SPIRE 2006), Glasgow, 11-13 October, pp. 98-109.

Beitzel, S.M., Jensen, E.C., Lewis, D.D., Chowdhury, A. and Frieder, O. (2007), "Automatic classification of web queries using very large unlabeled query logs", *ACM Transactions on Information Systems*, Vol. 25 No. 2.

Belkin, N.J. (1993), "Interaction with texts: information retrieval as information-seeking behavior", *Information retrieval ' 93. Von der Modellierung zur Anwendung, Proceedings of the 1st Conference of the Gesselschaft für Informatik Fachgruppe Information Retrieval, Universitaetsverlag Konstanz, Konstanz, Germany*, pp. 55-66.

Broder, A. (2002), "A taxonomy of web search", *SIGIR Forum*, Vol. 36 No. 2, pp. 3-10.

Cao, H., Hu, D.H., Shen, D., Jiang, D., Sun, J.-T., Chen, E. and Yang, Q. (2009), "Context-aware query classification", paper presented at the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Boston, MA, 19-23 July.

Caramel, E., Crawford, S. and Chen, H. (1992), "Browsing in hypertext: a cognitive study", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 22 No. 5, pp. 865-83.

Fujii, A. (2008), "Modeling anchor text and classifying queries to enhance web document retrieval", in Huai, J., Chen, R. and Hon, H.-W. (Eds), *Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 21-25 April*, pp. 337-46.

Han, J. and Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Francisco, CA.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York, NY.

Jansen, B.J. and Spink, A. (2005), "How are we searching the world wide web? A comparison of nine search engine transaction logs", *Information Processing & Management*, Vol. 42 No. 1, pp. 248-63.

Jansen, B.J. and Spink, A. (2007), "Sponsored search: is money a motivator for providing relevant results?", *IEEE Computer*, Vol. 40 No. 8, pp. 50-5.

Jansen, B.J., Booth, D. and Spink, A. (2008), "Determining the informational, navigational, and transactional intent of web queries", *Information Processing & Management*, Vol. 44 No. 3, pp. 1251-66.

Jansen, B.J., Spink, A. and Pedersen, J. (2005), "Trend analysis of AltaVista web searching", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 6, pp. 559-70.

Jansen, B.J., Zhang, M. and Spink, A. (2007), "Patterns and transitions of query reformulation during web searching", *International Journal of Web Information Systems*, Vol. 3 No. 4, pp. 328-40.

Kang, I. and Kim, G. (2003), "Query type classification for web document retrieval", paper presented at the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 28 July-1 August, pp. 64-71.

Korgaonkar, P., Silverblatt, R. and Girard, T. (2006), "Online retailing, product classifications, and consumer preferences", *Internet Research*, Vol. 16 No. 3, pp. 267-88.

Lee, U., Liu, Z. and Cho, J. (2005), "Automatic identification of user goals in web search", paper presented at the World Wide Web Conference, Chiba, Japan, 10-14 May, pp. 391-401.

Lu, H.-P. and Hsiao, K.-L. (2007), "Understanding intention to continuously share information on weblogs", *Internet Research*, Vol. 17 No. 4, pp. 345-61.

Marchionini, G. (1995), *Information Seeking in Electronic Environments*, Cambridge University Press, Cambridge.

Özmutlu, H.C., Çavdur, F. and Özmutlu, S. (2006), "Automatic new topic identification in search engine transaction logs", *Internet Research*, Vol. 16 No. 3, pp. 323-38.

Park, S., Bae, H. and Lee, J. (2005), "End user searching: a web log analysis of NAVER, a Korean web search engine", *Library & Information Science Research*, Vol. 27 No. 2, pp. 203-21.

Rose, D.E. and Levinson, D. (2004), "Understanding user goals in web search", paper presented at the World Wide Web Conference (WWW 2004), New York, NY, 17-22 May, pp. 13-19.

Rozanski, H.D., Bollman, G. and Lipman, M. (2001), "Seize the occasion! The seven-segment system for on-line marketing", *Strategy and Business*, Vol. 24 No. 3, pp. 42-51.

Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. (1999), "Analysis of a very large web search engine query log", *SIGIR Forum*, Vol. 33 No. 1, pp. 6-12.

Sullivan, D. (2006), Nielsen/NetRatings Search Engine Ratings, 23 February, available at: www.searchenginewatch.com/reports/netratings.html (accessed 1 June 2006).

Tann, C. and Sanderson, M. (2009), "Are web-based informational queries changing?", *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 6, pp. 1290-3.

About the authors

Ashish Kathuria is a Master's student at the Department of Electrical Engineering at the Pennsylvania State University, USA. His specific areas of interest are information search and web searching.

Bernard J. Jansen is an Associate Professor in the College of Information Sciences and Technology at The Pennsylvania State University, USA. He has contributed to more than 200 publications in the area of information technology and systems, with articles appearing in a multi-disciplinary range of journals and conferences. His specific areas of expertise are web searching, sponsored search, and personalization for information searching. He is co-author of the book, *Web Search: Public Searching of the Web*, co-editor of the book *Handbook of Weblog Analysis*, and author of *Understanding User-Web Interactions via Web Analytics*. Bernard J. Jansen is a member of the editorial boards of seven international journals. He has received several awards and honors, including an ACM Research Award and six application development awards, along with other writing, publishing, research, and leadership honors. Several agencies and corporations have supported his research. He is actively involved in teaching both undergraduate and graduate level courses, as well as mentoring students in a variety of research and educational efforts. Bernard J. Jansen is the corresponding author and can be contacted at: jjansen@ist.psu.edu

Carolyn Hafernik is a PhD student at the College of Information Sciences and Technology at the Pennsylvania State University, USA. Her specific areas of interest are information search and personalization.

Amanda Spink is Professor in the Faculty of Information Technology at the Queensland University of Technology and co-leader of the information science cluster. Her primary research includes: basic, applied, industry and interdisciplinary studies in information science; information behavior; cognitive information retrieval; and web retrieval; including relevance, feedback and multitasking models. Amanda Spink has published over 300 journal articles, refereed conference papers and book chapters, and five books. She is a member of numerous journal editorial boards including: *Information Processing and Management*, *Journal of Documentation*, *Journal of Information Systems Education*, and *Webology*.